Федеральное государственное образовательное бюджетное
учреждение высшего профессионального образования

«НАЦИОНАЛЬНЫЙ-ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Высшая школа бизнеса

Проектное задание
по дисциплине «Анализ и прогнозирование рыночных рисков»

Название проекта:
«Анализ и прогнозирование цен на золото на основе данных London Bullion Market Association с использованием машинного обучения»

3 курс, образовательная программа «Бизнес-информатика»

Работу выполнили:

Осипенкова Виолетта Андреевна,

Багрова Вера Дмитриевна,

Новиков Виталий Валерьевич

Москва 2023

Оглавление

Актуальность **3**
Глава I **3**
Данные 3
Методы вычисления 4
Глава II **5**
Графики 5
Box-plot 5
STL анализ 7
Выбросы 11
Автокорреляция 12
Частная автокорреляция 13
Предсказания 14
Попытки предсказаний 14
Предсказание AutoRegressivePipeline 15
Предсказание Pipeline 16
Будущие цены 17
Выводы **18**
Ссылки и источники **19**

Актуальность

В рамках анализа было выбрано цена в долларах за тройскую унцию золота по данным London Bullion Market Association (далее LBMA) (источник https://investfunds.ru/indexes/354/)

Актуальность проекта заключается в том, что в мае 2023 года рассматриваемый индекс достиг рекордно большого значения с 2020 года. При этом за десятилетнюю историю у индекса было всего три момента, когда цена поднималась выше 2000 долларов. Сейчас мы наблюдаем важный момент в рамках ценовой оценки золота в мире, поэтому проект несет в себе большую ценность.

Глава I

Данные

LBMA Gold Price это ключевой индикатор для определения цены на золото, широко используемый производителями, покупателями, инвесторами и центральными банками по всему миру. Цены устанавливаются в долларах за тройскую унцию в 15:00 по Лондонскому времени.

рис.1 исходные данные

График цены золота прерывистый, так как отсутствуют некоторые данные. Эти данные отсутствуют по выходным дням, поэтому заполним пропуски через скользящее среднее с окном 30, которое сглаживает данные и уменьшает воздействие шума. Скользящее среднее - это способ заполнения пропущенных значений путем усреднения значения предыдущих и последующих точек. В нашем случае скользящее среднее за 30 дней усредняет значения цены за 15 дней до, 15 дней после.

С бизнесовой точки зрения отсутствие данных можно объяснить тем, что биржа не работает в выходные дни, поэтому фиксируется только значение в пятницу, а затем в понедельник.

рис.2 Данные с заполненными пропусками

Как видно из графика, после заполнения пропусков цена золота ведет себя более плавно и имеет более развернутые экстремумы. Это объясняется тем, что скользящее среднее помогает сгладить некоторые выбросы в данных, сглаживая колебания цен в определенный период времени.

Методы вычисления

Для прогнозирования мы пользовались открытой библиотекой ETNA, разработанной компанией Тинькофф. Работа с библиотекой состоит из нескольких этапов:

1) Подготовка и валидация данных

Для работы с данными в библиотеке ETNA доступен TSDataset-класс. С помощью него можно привести различные ряды к единому формату, восстановить потерянные данные по частотности, а также установить связь прогнозирования данных с дополнительными данными.

2) Предварительный анализ данных (EDA)

Для определения и понимания структуры и особенности прогнозируемых рядов, существуют методы EDA. Они позволяют построить статистики по данным, оценить автокорреляцию, обнаружить выбросы.

3) Построение пайплайна прогнозирования

По результатам EDA можно понять, какие признаки выделять из данных, как нужно обработать ряды для дальнейшей работы. Например, вычесть тренд или прологарифмировать.

4) Построение прогноза и валидация

Чтобы проверить, насколько хорошо представленный пайплайн будет работать для данных рядов, можно запустить тестирование на исторических данных.

Глава II

Графики

Box-plot

Прежде чем приступать к анализу существующих графиков, обратимся к визуализации цен по дням недели в виде бокс-плотов.

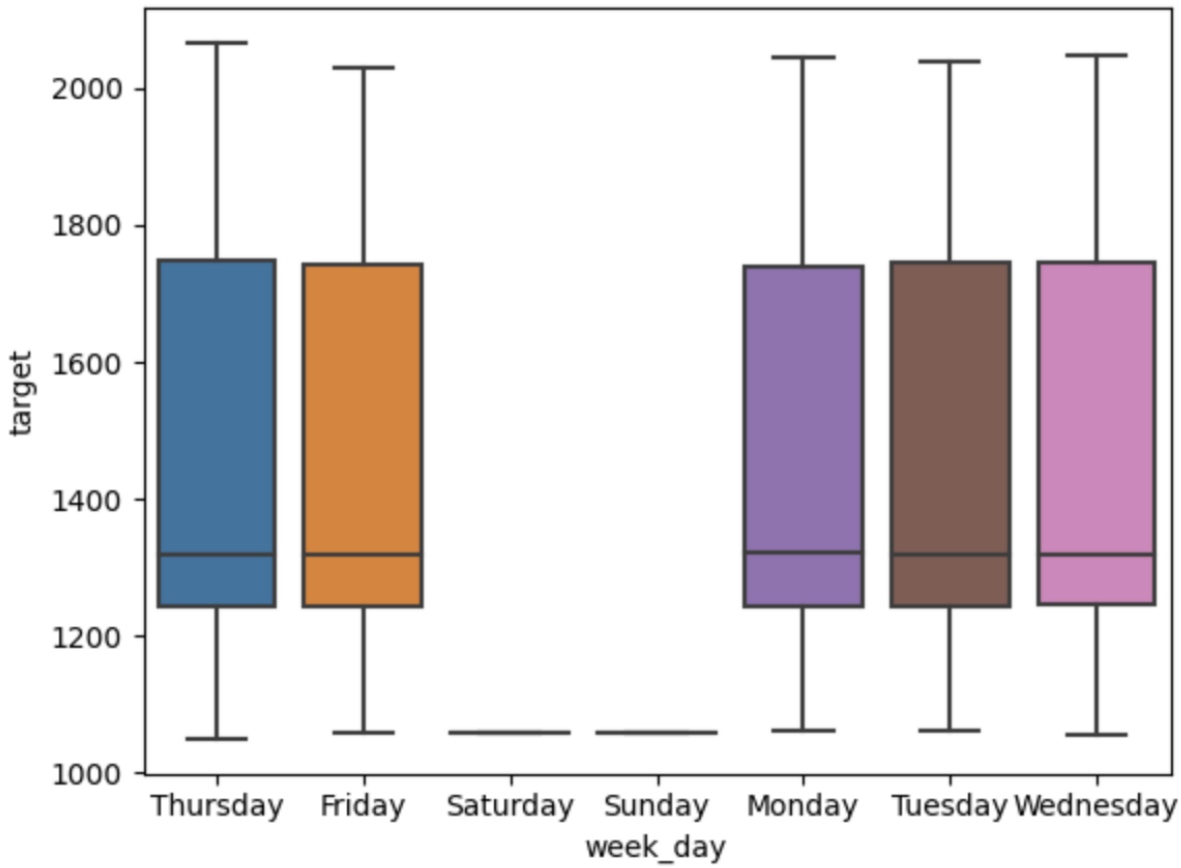

рис. 3 box-plots на каждый день недели кроме выходных

Первым делом упомянем тот факт, что здесь нет данных по субботам и воскресеньям? потому что нам интересно посмотреть на исходные данные. Нетрудно догадаться, что медиана у этих графиков примерно на одном уровне, ведь в действительности количество замеров каждого дня одинаковое при этом довольно большое, поэтому медиана, 25 и 75 перцентили находятся примерно на одном уровне. Однако можем заметить, что максимальное значение четверга выше максимальных значений остальных дней недели, что может значить, что наивысшая точка графика на выбранном периоде приходится именно на четверг, при этом эта цена продержалась недолго (на следующий день упала).

Далее будем анализировать данные с заполненными пропусками в выходные. О методе заполнения мы писали ранее.

STL анализ

Алгоритм STL сглаживает временные ряды с помощью LOESS в двух циклах; внутренний цикл выполняет итерацию между сезонным сглаживанием и сглаживанием тренда, а внешний цикл минимизирует влияние выбросов. В ходе внутреннего цикла сезонный компонент вычисляется первым и удаляется для вычисления компонента тренда. Остаток вычисляется путем вычитания сезонного и трендового компонентов из временных рядов.

Попробуем сделать STL анализ на наших данных. При этом будем использовать несколько значений для параметра period, чтобы попробовать найти наилучший вариант.

(рисунки 4, 5 и 6 приведены ниже по тексту)

Заметим, что каждое из разложений имеет довольно много шума, что характерно для данных такого типа (биржевые ежедневные цены). В каждом из трех случаев мы видим похожий тренд, однако при увеличении параметра period тренд становится чуть более сглаженным. Общий тренд демонстрирует нам, что до 2019 года цена оставалась в пределах 1100 - 1400 долларов, а с 2019 по 2021 поднималась до нового плато 1700 - 1900 долларов. С начала 2023 года мы видим восходящий тренд, который может являться выходом на новое плато. При этом ни в одном из случаев нельзя утверждать, что нам удалось выделить сезонность, ведь ни один график не является четко выраженной синусоидой. Однако при period = 365 мы видим повторяющиеся паттерны поведения: перед началом нового года цена снижается примерно на 75 долларов от тренда (особенно это видно с 2014 по 2018, но и позже подобное поведение прослеживается, хоть и с большими выбросами). Сейчас цена держится около 1950 долларов за тройскую унцию, при этом мы видим что график достиг сезонного плато и можем предполагать, что летом цена будет оставаться в пределах 1850 - 2100 долларов, а после может начать постепенно снижаться, чтобы к концу года достичь минус 50-100 к тренду.

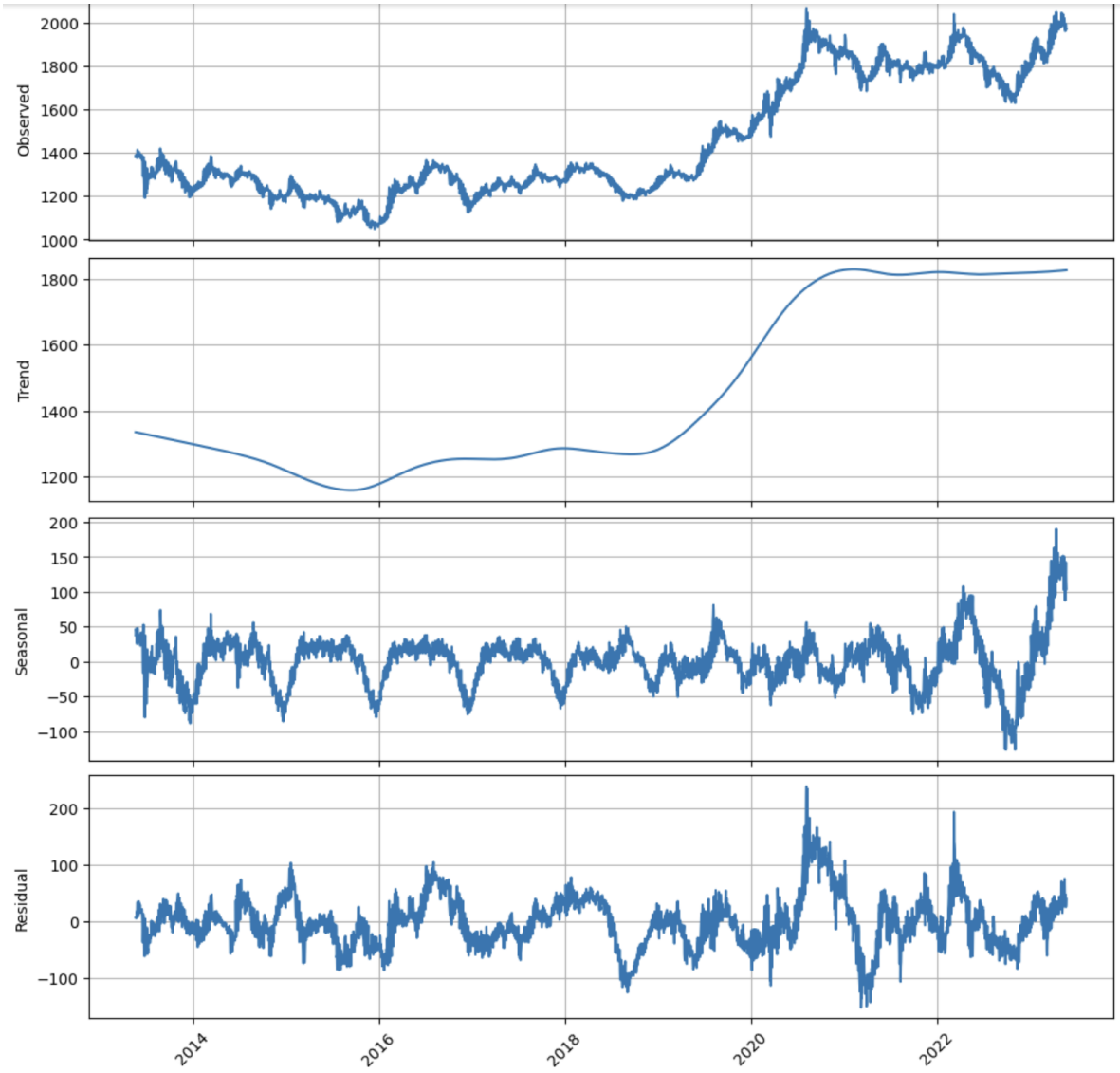

Рис. 4 STL-анализ при period = 365

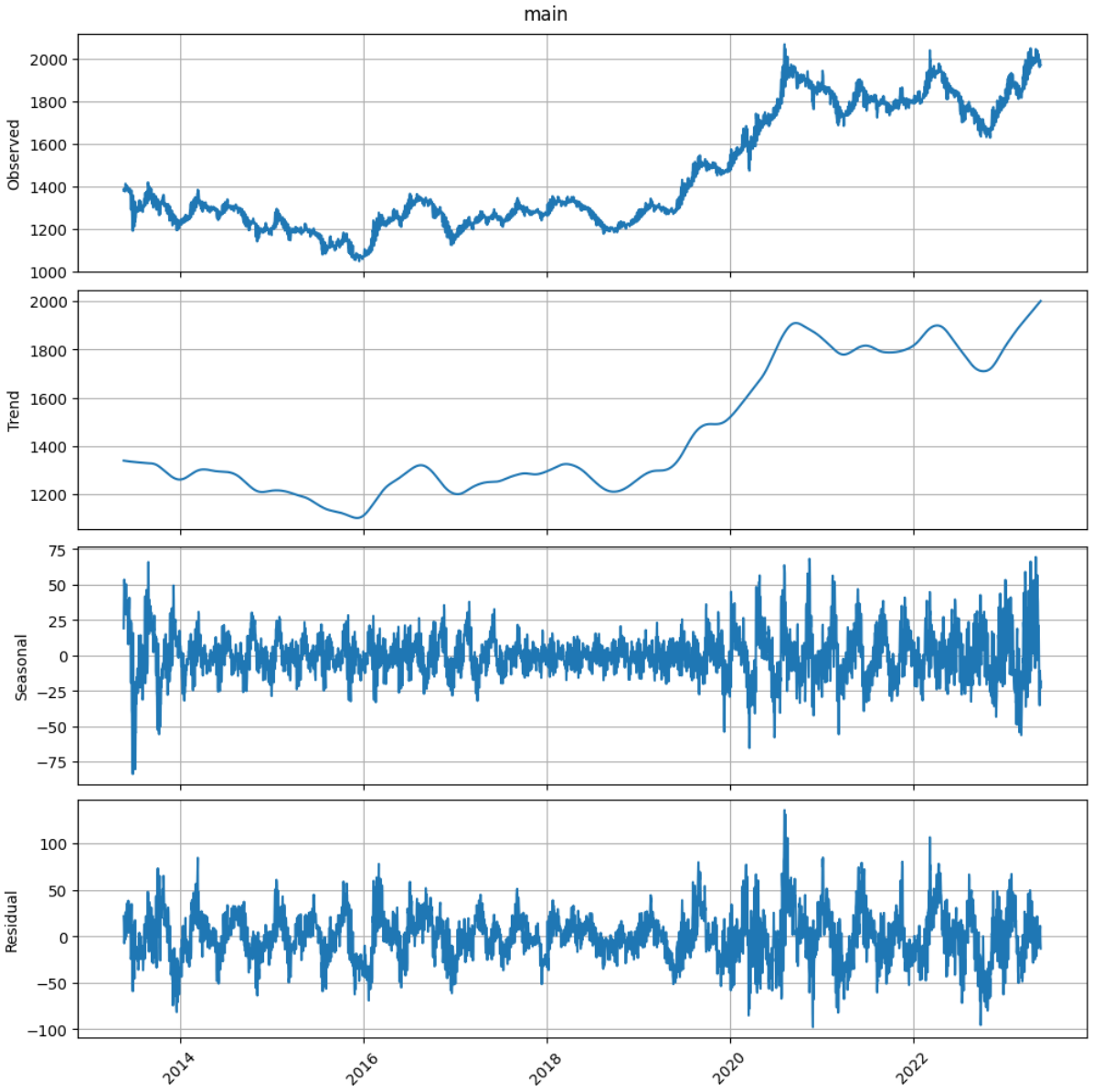

Рис. 5 STL-анализ при period = 100

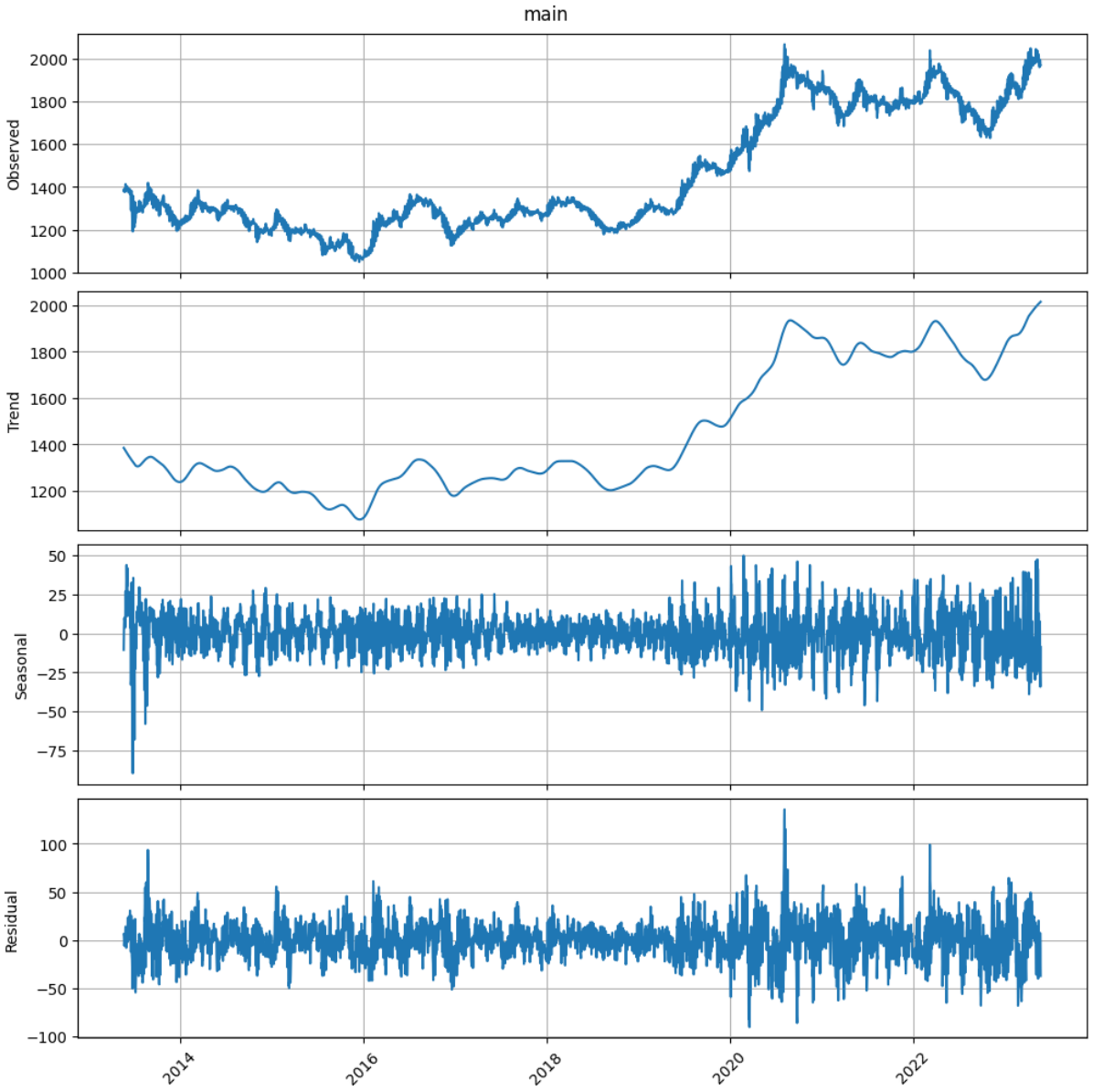

Рис. 6 STL-анализ при period = 50

Выбросы

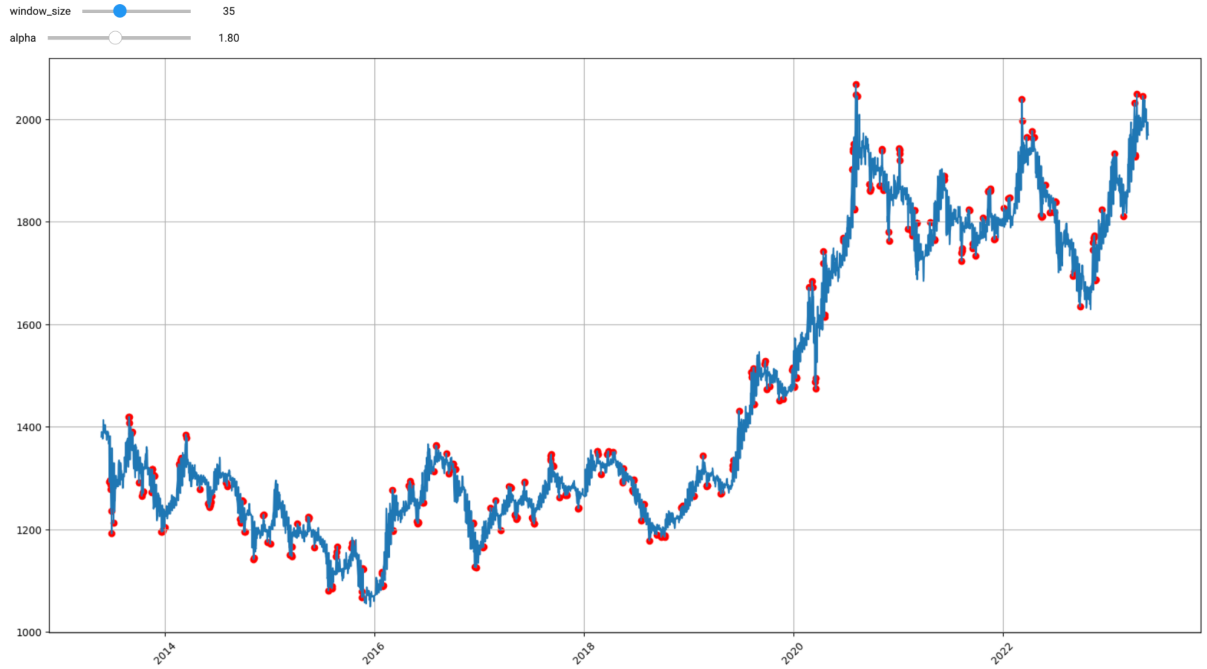

Рис. 7 - Выбросы по медиане с окном 35 и alpha = 1.8

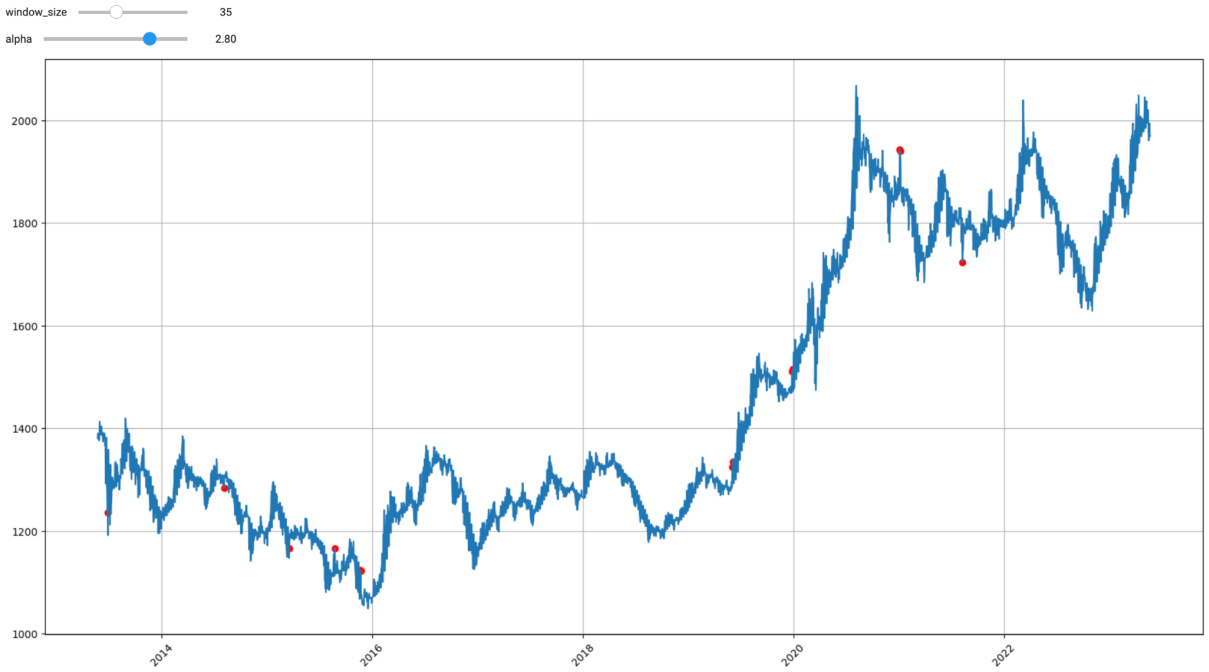

Рис. 8 - Выбросы по медиане с окном 35 и alpha = 2.8

Как и упоминали ранее, для данных характерно большое количество выбросов, поэтому при анализе выбросом мы видим большое количество красных точек (рис. 7), которые алгоритм определяет как значимые выбросы, но мы видим что в большинстве они являются свойственными

биржевым индексам спадам или повышениям цены. Чтобы уменьшить количество красных точек, которых алгоритм считает выбросами, мы можем увеличить параметр alpha, который регулирует отклонение от среднего. На рисунке 8 мы видим уже гораздо меньшее количество выбросов. Яркие примеры выбросов мы можем наблюдать на 4 января 2022 года, когда цена довольно резко выросла в рамках выбранного периода (35 дней), и 10 августа, когда цена резко снижалась.

Автокорреляция

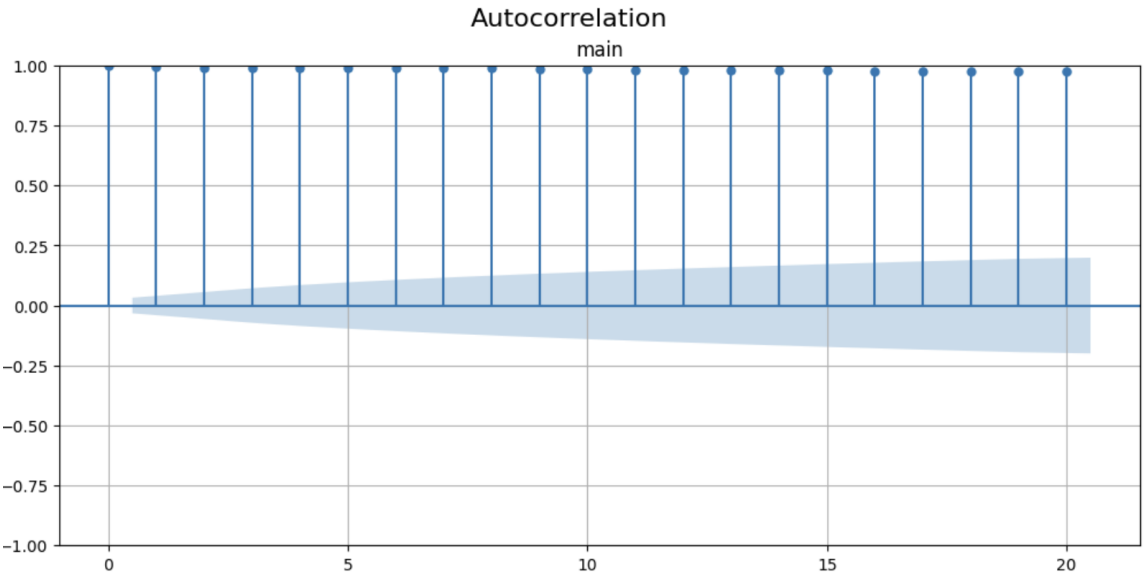

Рис. 9 - автокорреляция

Функция автокорреляции (АКФ). При лаге k это корреляция между рядами значений, отстоящих друг от друга на k интервалов, считая значения интервалов в промежутке.

По графику ACF можно определить необходимую величину лага для моделей скользящего среднего
В данном случае автокорреляция не дала результатов, поэтому обратимся к частной автогоррелции

Частная автокорреляция

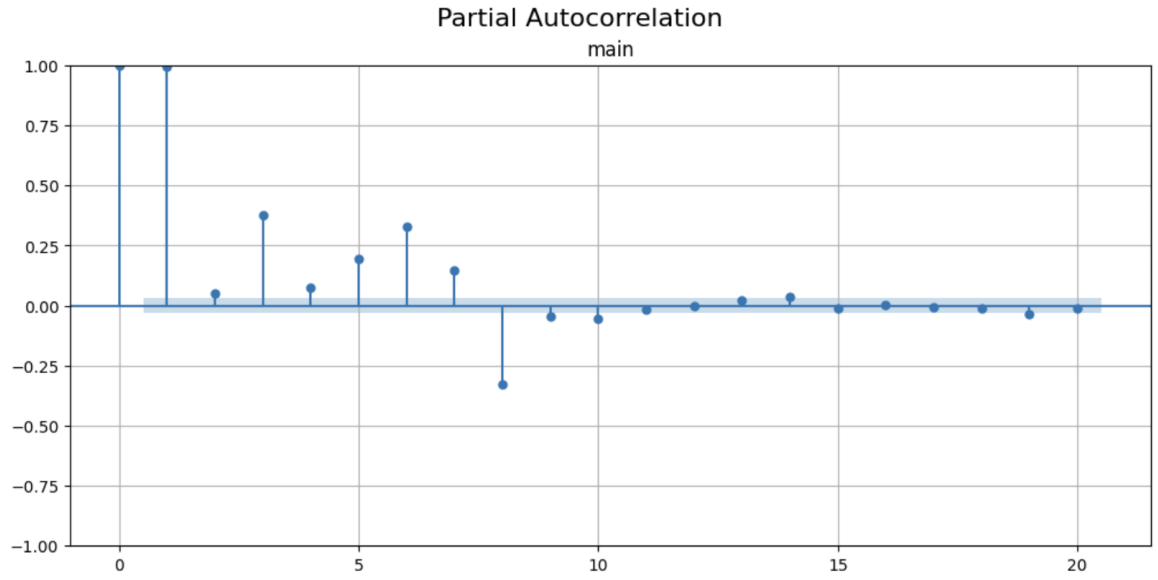

Рис. 10 - частная автокорреляция (lags = 20)

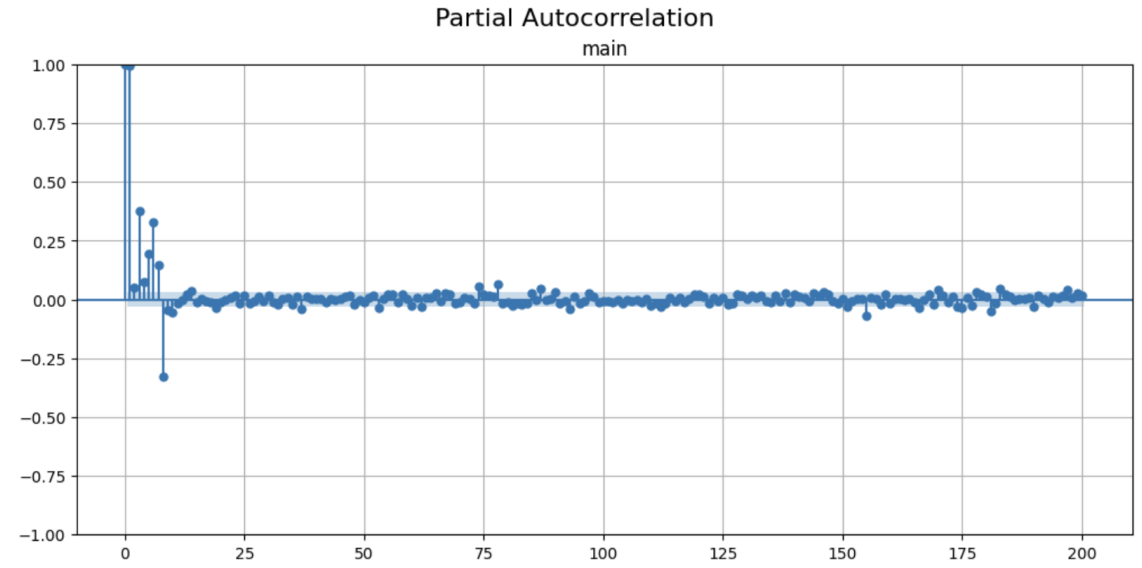

Рис. 11 - частная автокорреляция (lags = 200)

Функция частной автокорреляции (ЧАКФ). При лаге k это корреляция между рядами значений, отстоящих друг от друга на k интервалов.

По графику PACF можно определить необходимую величину лага для авторегрессионных моделей

В нашем случае видно, что в качестве лага достаточно брать 10 предыдущих значений, ведь все остальные имеют околонулевое значение функции частной автокорреляции.

Предсказания

Попытки предсказаний

Используется пайплайн AutoRegressivePipeline, который итеративно строит предсказание на step шагов вперед, после чего использует предсказанные значения как признаки для следующих шагов.

Плюсы пайплайна:

- Позволяет использовать лаги меньше значения параметра HORIZON
- Демонстрирует стабильность в случае рядов с высоким соотношением сигнал/шум

Минусы пайплайна:

- Может работать медленно в случае малых значений step, поскольку требует пересчета признаков $\lceil horizon \,/\, step \rceil$ раз
- Может быть неточным на больших горизонтах из-за потенциального накопления ошибки

Были проделаны первые попытки предсказания на обученной модели (рис. 12)

Рис. 12 - Предсказание AutoRegressivePipeline

Как видим стандартные входные параметры выдали нам даже не близкий к реальности прогноз. При этом модель определила странную какую-то сезонность и восходящий тренд. Ошибка такой модели крайне велика, поэтому попробуем изменить входные параметры.

Предсказание AutoRegressivePipeline

Попробуем добавить stl разложение на основе ARIMA, чтобы исправить ошибочное определение сезонностей.

ARIMA (autoregressive integrated moving average) - авторегрессионное интегрированное скользящее среднее является обобщением модели авторегрессионного скользящего среднего. Обычно модель упоминается, как ARIMA(p,d,q), где p,d и q — целые неотрицательные числа, характеризующие порядок для частей модели (соответственно авторегрессионной, интегрированной и скользящего среднего) [4]

Идея моделей семейства ARIMA: приведение ряда к стационарному и применение регрессии

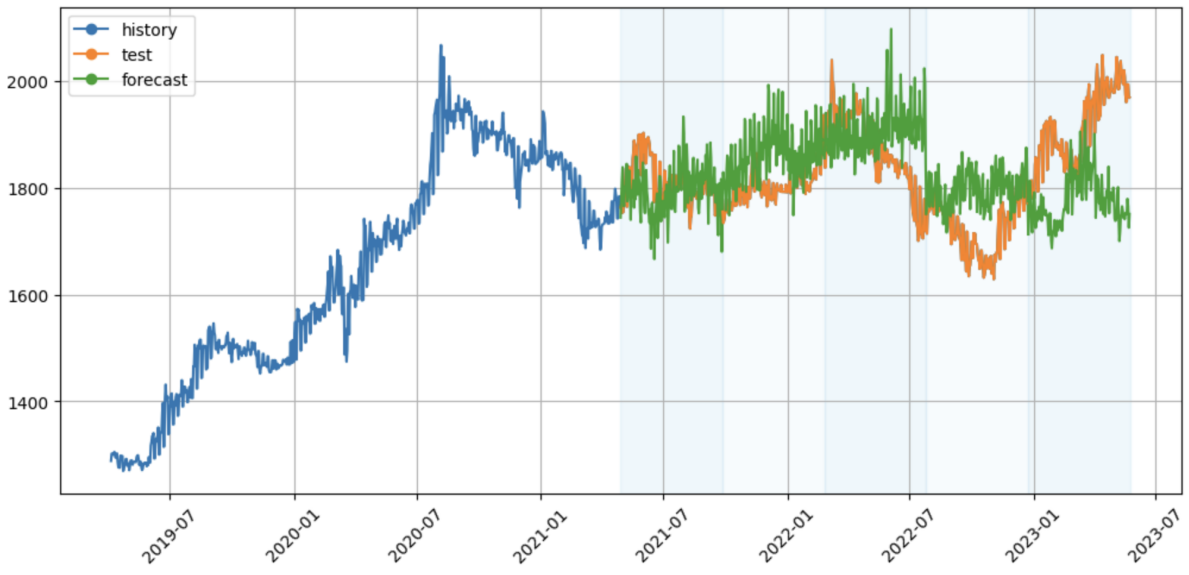

Рис. 13 - Предсказание AutoRegressivePipeline с дополнением ARIMA

Такое предсказание уже больше подходит, особенно на начальных этапах (до лета 2022 года). После середины 2022 года предсказание имеет уже больше ошибок, однако это связано с накопительной ошибкой выбранной

архитектуры, ведь мы используем предсказанные значения как признаки и как следствие ошибка сдвигается.

Предсказание Pipeline

Попробуем применить другую архитектуру, основанную на той же модели Pipeline реализует версию прямой стратегии, при которой всего одна модель обучается предсказывать все будущие значения. Это подразумевают следующее:

Плюсы:

- Самый быстрый метод - как на обучении, так и на инференсе

Минусы:

- Pipeline не позволяет использовать лаги со значениями меньше параметра horizon
- При использовании лагов может начать терять в качестве при увеличении горизонта, так как вслед за этим возрастает ограничение снизу на длину лагов

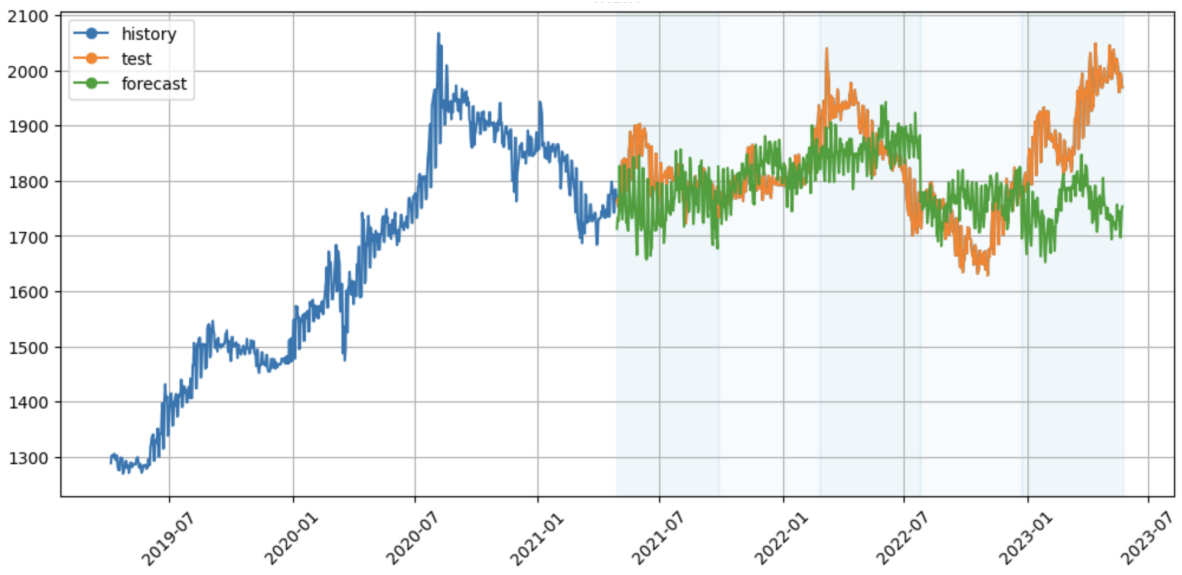

Рис. 14 - Предсказание Pipeline

Эта архитектура тоже дала довольно хорошие результаты (рис. 14). Мы видим хорошее сходство графиков до начала 2023 года. Хочется подчеркнуть хорошее предсказание падения цены в июле 2022 года, которое практически совпало с действительностью (оранжевый график).

Будущие цены

Применим нашу модель для предсказания цен на год вперед

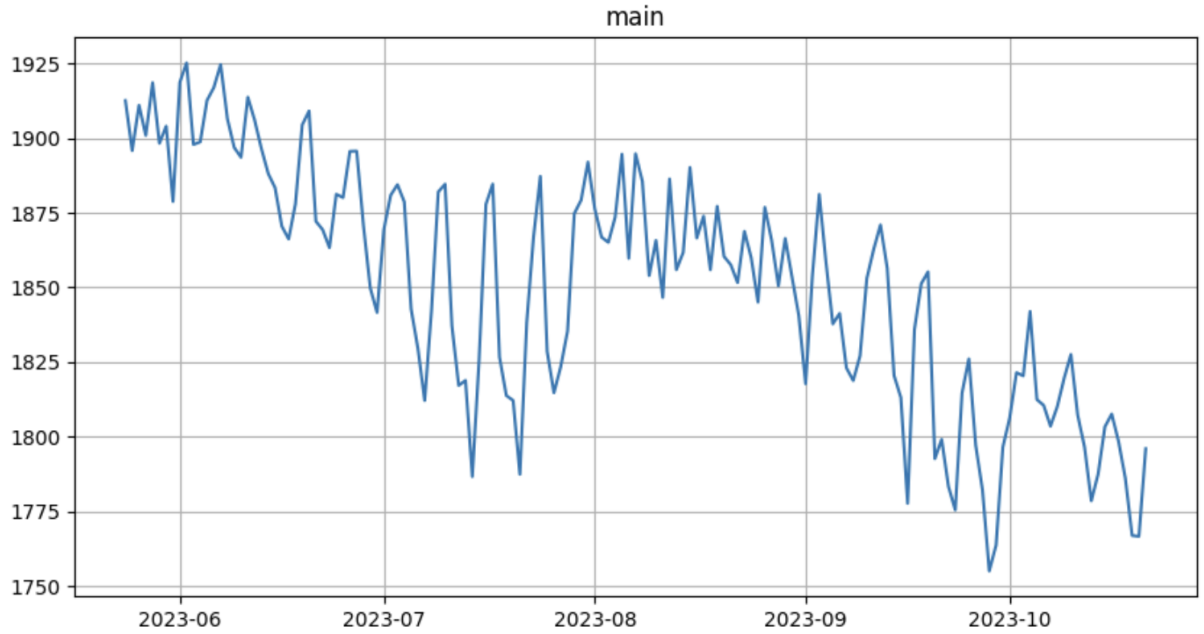

Рис. 15 - Предсказание модели на цены с июня по ноябрь 2023 года

Первый рассматриваемый нами период предсказаний будет на ближайшие 6 месяцев. Можем заметить, что модель предлагает нам вариант развития событий, когда цены не поднимаются выше 1925 долларов в ближайшие 6 месяцев, более того они будут долгое время колебаться в пределах 1800 - 1900 долларов. Ранее мы упоминали, что предварительный анализ наводил на мысли, что цена достигла нового плато, которое как раз предсказано моделью. Также хочется отметить, что модель смогла уловить наше наблюдение о снижении цены ближе к концу календарного года, за которым чаще всего идет новый подъем (рис. 16).

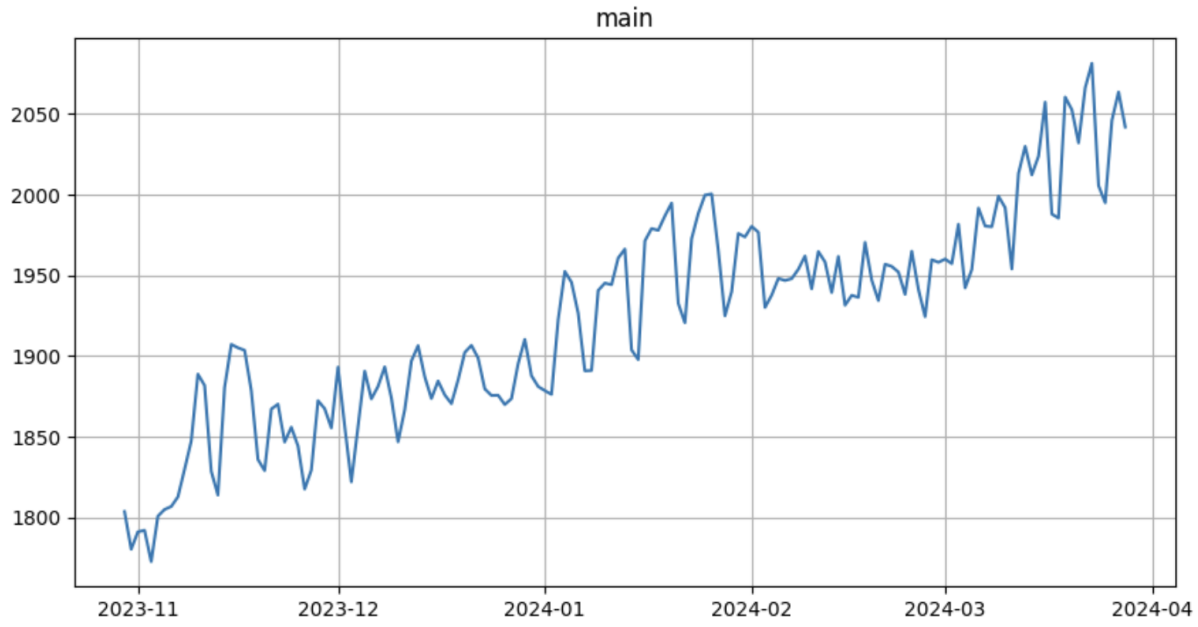

Рис. 16 - Предсказание модели на цены с ноября 2023 года по апрель 2024 года

Это уже следующие 6 месяцев прогноза нашей модели. Тут мы уже видим восходящую часть сезонности, что в комплексе с восходящим трендом дает предсказание о достижении цены в 2000 долларов за тройскую унцию золота уже в конце января, а далее будут значения, превосходящие 2050 к концу периода предсказаний.

Выводы

Таким образом, на основе данных о цене тройской унции золота на лондонской бирже нам удалось проанализировать тренды, сезонности, периоды и выбросы. Помимо этого были реализованы модели на основе разных архитектур прогнозирования временных рядов, среди которых удалось выбрать наиболее подходящую для построения прогноза цен на следующий год. Был построен прогноз цен по рассматриваемому индексу до апреля 2024 года, который подчиняется подчеркнутым нами трендам, сезонностям и паттернам поведения цен, что говорит о возможной точности прогнозов.

Ссылки и источники

1) https://investfunds.ru/indexes/354/
2) https://etna-docs.netlify.app/api/etna.pipeline.autoregressive_pipeline.autoregressive pipeline
3) https://etna-docs.netlify.app/api/etna.pipeline.pipeline.pipeline#etna.pipeline.pipeline.Pipeline
4) https://etna-docs.netlify.app/analysis.html